Utmel: Post-ChatGPT Era, How AI Chips Can Succeed

ChatGPT, a new artificial intelligence chatbot, is considered to be "starting a new AI revolution" with its unprecedented momentum.

HONGKONG, February 28, 2023 /EINPresswire.com/ -- There is no doubt that ChatGPT is far hotter than its predecessor AlphaGo, and has clearly illuminated the prospect of large model applications like a torch. However, when the iPhone was introduced, many outside observers were still accustomed to linear extrapolation of the existing vendor landscape, believing that the "new world" explored by Apple would ultimately remain in Nokia's pocket.

After the "iPhone moment of artificial intelligence", new opportunities in the AI industry are not rightfully promised to the old players.

It is worth pointing out that many teams face the dual constraints of engineering capability and economic cost to complete a large model from 0 to 1 that replicates and even surpasses the actual ChatGPT experience.

In terms of engineering capability, the number of large language model (LLM) parameters has moved from billions to hundreds of billions in just a few years, and the demand for arithmetic power has far exceeded the iterative rate of processor performance; distributed parallel computing has thus entered "deep water", and embarrassingly parallel methods have hit the ceiling and must be developed. The distributed parallel computing method has hit the ceiling, and more complex scheduling methods between sub-computing tasks and processors must be developed, and related talents are currently scarce.

In terms of economic cost, the "mother" GPT-3 model of ChatGPT, for example, is said to have a theoretical minimum cost of $4.6 million for a single training session using NVIDIA V100 GPU clusters, without considering the tuning and scheduling, while the supercomputing system built by Microsoft specifically for its developer OpenAI is said to have The supercomputing system built by Microsoft specifically for its developer OpenAI is said to have more than 280,000 CPU cores and 10,000 GPUs, with overall performance reaching the top five of the global TOP500 supercomputers list by 2020. If this propaganda is basically true, it is equivalent to OpenAI using a complete Tianhe 2 supercomputing system specifically to support its model training, which is undoubtedly too extravagant for most companies.

In this view, the current end-to-end vertical integration model of AI vendors from data and algorithms to AI chips, hardware terminals, and project implementation may move toward a clearer professional division of labor in the future, with a few tech giants with giant computing clusters conducting large-scale pre-training model iterations in the cloud and opening interfaces to downstream vendors and developers, and the industry chain downstream based on domain-specific knowledge on the side and end side to complete models at lower cost and shorter cycle time. The downstream of the industry chain will complete model fine-tuning at the edge and end side based on domain-specific knowledge at a lower cost and shorter cycle to achieve highly available delivery for vertical application scenarios.

In the face of the emerging "paradigm shift", for most of the new and old vendors, it is more important to think deeply about how to tap the business value of specific scenarios than to send news that "we have a similar model under development".

Utmel pointed out that in reality, there is no doubt that the platform giants have advantages in terms of R&D investment and team investment. However, with the extension of ChatGPT's application, both upstream and downstream enterprises will contribute to the key links of the industry chain. For downstream manufacturers, exploring the market application of GPT-like products and realizing the commercial value of the technology are the key concerns of enterprises, whether it is product tools or product solutions, mining potential scenarios, conducting technology integration, and outputting holistic and result-oriented practical solutions are the real issues.

Utmel said that ChatGPT is now presented in the form of a text interaction robot, this kind of text generation actually has a lot of application space, for example, in the direction of intelligent customer service, now the intelligent customer service is retrieval AI, but generative AI initiative is higher more affinity and effectiveness. Whether it is a search engine, e-commerce customer service or AI-assisted generation, ChatGPT has a strong capacity for application-level innovation.

Utmel predicts that ChatGPT is a suitable ground for landing in creative industries that need to be based on certain background knowledge, as well as scenarios that just need AIGC and industries with SOPs (Standard Operating Procedures), such as smart writing, smart customer service, document management, code generation, and even game NPCs. Big model technology can be applied to achieve the integration of scenarios by strengthening context understanding ability, thinking chain reasoning, and enhancing instruction learning. For example, in the meeting scenario, based on thousands of words of meeting records, the tool can quickly organize the meeting outline and focus and clearly list the to-do items according to the demand instructions.

Further specific to the field of voice interaction, Utmel believes that the future to voice dialogue robots to go to progress, strengthen the application of multimodal interaction technology of voice, text, image and other deep integration, to cope with the changes of complex scenes. All these leave rooms for thinking, application, and exploration.

As the number of large models evolves from tens of billions, hundreds of billions to trillions, the new AI industry competition will further focus on the arithmetic link, and at the same time, the difference in demand for upstream and downstream reasoning and training workloads will become more and more significant, which also brings new traction to the evolution of chip technology.

Utmel analyzes that GPT3.5 behind ChatGPT is a disruptive underlying technology whose stunning application effects are built on the basis of a huge corpus and ultra-large-scale AI arithmetic power. As the application scenarios evolve, the core technology will accelerate, including the complexity of AI models will continue to evolve, which will generate a well of demand for computing power. The future continues to be optimistic about the development of general-purpose GPU architecture training products, its versatility, compatibility, and ecological maturity are still the main support for the construction of AI algorithms and applications in the coming period.

Compared with the cloud-side demand for versatility and scalability, inference chips face different needs on the side-side and end-side.

Utmel said that AI-specific chips can achieve site-specific low energy consumption and high arithmetic power on the end-side that focuses on scenarios. From a customization point of view, ASIC-specific AI chips have more advantages from an efficiency perspective, and with the popularity and application of large models, they can improve the cost performance of related chip products.

At the same time, due to the dramatic increase in the volume of data, whether in the cloud side or end side, the privacy protection of data has also put forward higher requirements. In order to ensure the benign development of AI, it is necessary to go to embed some corresponding restrictive means and rule constraints. AI-based applications and other derived tool-level products are in urgent need of regulation in terms of privacy and security, intellectual property risks, etc.

Sophia
Utmel Electronic CO.,LTD
+86 13189752889
email us here
Visit us on social media:
Facebook
Twitter
LinkedIn

try to be careful about weeding out false and misleading content. As a user, if you see something we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.